systematic reviews (EFSA, 2010). Then, the risk of bias is assessed based on the factors described in section 6.2 for a WoE assessment, namely: study design and conduct, population, exposure assessment, outcome assessment, confounder control, statistical analysis and reporting of results. Those studies categorised as of low reliability will be considered unacceptable for risk assessment. The remaining studies will be weighted and used for hazard identification.

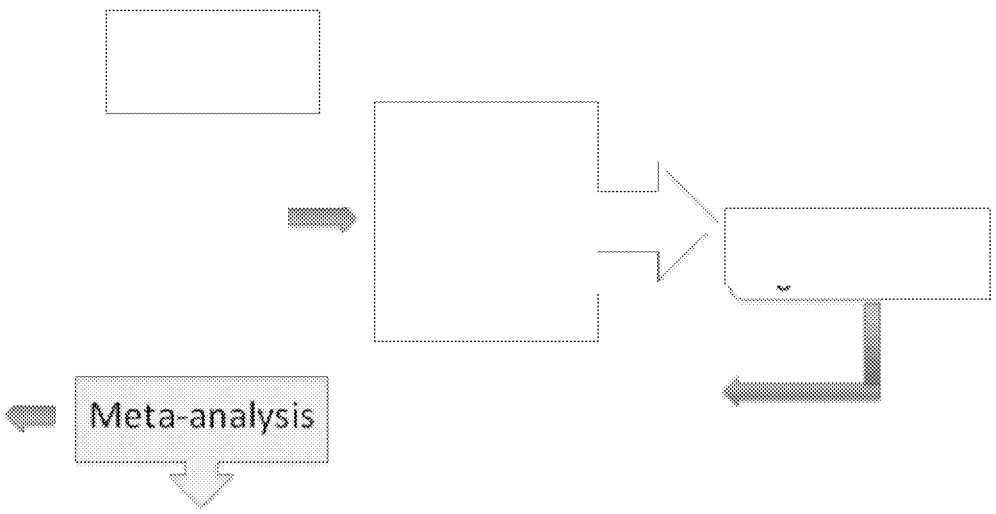

Figure 3: Methodology for utilization of human data for risk assessment.

If quantitative data are available, a meta-analysis can be conducted to create summary data and to improve the statistical power and precision of risk estimates (OR, RR) by combining the results of all individual studies available or meeting the selection criteria. As meta-analyses determine the size of association averaged over the considered studies, they provide a stronger basis for hazard identification. Moreover, under certain circumstances, there is the possibility to move towards risk characterization metrics because these measured differences in health outcomes (OR, RR) can be converted to dose-response relationships (Nachman et al., 2011). Although quite unusual in practice, this would allow for the identification of critical effects in humans and/or setting reference values without the need of using animal extrapolation.

Since heterogeneity is common in meta-analyses, there is a need to assess which studies could be combined quantitatively. Heterogeneity can be genuine, representing diverse effects in different subgroups, or might represent presence of bias. If heterogeneity is high (I^2 greater than 50%), individual studies should not be combined to obtain a summary measure because of the high risk of aggregating bias from different sources. Sources of heterogeneity should be explored through sensitivity analysis and/or meta-regression. Furthermore, the presence of diverse biases in the meta-analysis should be examined, such as small study effects, publication bias and excess significance bias. It is important to find models that adequately describe the effect-size distribution of the underlying studied populations.

6.3.4. Pooling data from similar epidemiologic studies for potential dose-response modelling

As in other fields of research, findings from a single epidemiological study merit verification through replication. When the number of replications is abundant, it may be worthwhile to assess the entire set of replicate epidemiological studies through a meta-analysis and ascertain whether, for key outcomes, findings are consistent across studies. Such an approach will provide more robust conclusions about the existence of cause-effect relationships.

Once a hazard has been identified, the next step in risk assessment is to conduct a dose–response assessment to estimate the risk of the adverse effect at different levels of exposure and/or the concentration level below which no appreciable adverse health effect can be assumed for a given population.

However, this step requires fully quantitative (or at least semi-quantitative) exposure data at individual level. Summary estimates resulting from quantitative synthesis would be more informative for risk assessment if they present OR for a given change in the continuous variable of exposure (or per a given percentile change in exposure) as this allows for relative comparisons across studies and could be of help to derive health-based reference values. Only within such a framework can data from human studies with similar designs be merged to gain enough power to model proper dose-response curves (Greenland and Longnecker, 1992; Orsini et al., 2012).

Conversely, meta-analytical approaches may be of limited value if a combined OR is calculated based on meta-analyses interpreting exposure as a 'yes' or a 'no' because exposures are not necessarily to active ingredients in the same proportion in all studies included. Even though in these cases meta-analyses may consistently find an increased risk associated with pesticide exposure, for risk assessment the exposure needs to characterise the effect of specific pesticide classes or even better individual pesticides (Hernández et al., 2016).

This approach would allow points of departure to be identified (e.g., benchmark doses -BMD-) and would be relevant for the integration of epidemiological studies into quantitative risk assessment. Although BMD modelling is currently used for analysing dose-response data from experimental studies, it is possible to apply this approach to data from observational epidemiological studies. The EFSA Scientific Committee confirmed that the BMD approach is a scientifically more advanced method compared to the NOAEL approach for deriving a Reference Point, since it makes extended use of the dose-response data from experimental and epidemiological studies to better characterise and quantify potential risks. This approach, in principle, can be applicable to human data (EFSA 2017b).

Dose-response data from observational epidemiological studies may differ from typical animal toxicity data in several respects and these differences are relevant to BMD calculations. Exposure data often do not fall into a small number of well-defined dosage groups. Unlike most experimental studies, observational studies may not include an unexposed control group, because all individuals may be exposed to some extent to a chemical contaminant. In this case, the BMD approach still applies since fitting a dose-response curve does not necessarily require observations at zero exposure. However, the response at zero exposure would then need to be estimated by low-dose extrapolation. Hence the BMD derived from epidemiological data can be strongly model-dependent (Budtz-Jørgensen et al., 2001).

7. Integrating the diverse streams of evidence: human (epidemiology and vigilance data) and experimental information

This chapter first considers in 7.1 the different nature of the main streams of evidence, i.e. originating either from experimental studies or from epidemiological studies. The approach used is that recommended by the Scientific Committee Opinion on WoE (2017b), which distinguishes 3 successive phases to assess and integrate these different streams of information: reliability, relevance and

consistency. The first step, consists in the assessment of the reliability of individual studies be they epidemiological (addressed in chapter 6) or experimental. Then, the relevance (strength of evidence) of one or more studies found to be reliable is assessed using principles of epidemiology (addressed in chapter 6) and toxicology. Next, section 7.2 considers how to bring together different streams of relevant information from epidemiological and experimental studies, which is considered in a WoE approach, to assess consistency and biological plausibility for humans.

7.1. Sources and nature of the different streams of evidence Comparison of experimental and epidemiological approaches

In the regulatory risk assessment of pesticides, the information on the toxic effects is based on the results of a full set of experiments as required by Regulation (EC) 283/2013 and 284/2013, and conducted according to OECD guidelines. They are carried out *in vivo* or *in vitro.* A number of categories are established for rating the reliability of each stream of evidence according to the EFSA peer review of active substances: acceptable, supplementary and unacceptable. The data quality and reliability of in vivo or in vitro toxicity studies should be assessed using evaluation methods that better provide more structured support for determining a study's adequacy for hazard and risk assessments. Animal (in vivo) studies conducted according to standardized test guidelines and good laboratory practices (e.g. OECD TG) are by default attributed higher reliability than other research studies. Notwithstanding, since there is no evidence that studies conducted under such framework have a lower risk of bias (Vandenberg et al., 2016), evidence from all relevant studies, both GLP and non-GLP, should also be considered and weighted. Besides, the internal validity of *in vitro* toxicity studies should be evaluated as well to provide a better support for determining a study's adequacy for hazard and risk assessments. *In silico* modelling can be used to derive structure-activity relationships (SAR) and to complement current toxicity tests for the identification and characterization of the mode or mechanisms of action of the active substance in humans. These alternative toxicity testing approaches could be helpful in the absence of animal data, e.g. to screen for potential neurodevelopmental or endocrine disruption effects of pesticides, and to increase confidence in animal testing.

Besides toxicity data on the active substance, such data may also be required on metabolites or residues if human exposure may occur through the diet or drinking water. Results from these studies are then considered in relation to expected human exposures estimated through food consumption and other sources of exposure. The strength of this approach is that experimental studies in laboratory animals are controlled studies where confounding is eliminated by design, which is not the case with epidemiological studies. Animals used in regulatory studies are, however, typically inbred, genetically homogeneous and due to the controlled environment they lack the full range of quantitative and qualitative chemical susceptibility profiles.

Many experimental models do not capture complex multifactorial diseases making animal-to-human extrapolation subject to considerable uncertainty. Current risk assessment is therefore by its nature predictive and may be insufficient because it is chemical-specific and humans are exposed to a large number of chemicals from environmental, dietary and occupational sources or because of different toxicokinetic differences. In recognition of the uncertain nature of animal-to-human extrapolation the regulatory risk assessment advice does not just consider the relevant point(s) of departure (NOAEL, LOAEL or BMDL) that have been identified as safe but lowers these values using uncertainty factors (UFs) to propose safe reference dose values, either for acute or chronic toxicity.

In contrast, epidemiological studies examine associations between actual exposures in humans with disease. Epidemiological studies incorporate the true (or estimated) range of population exposures, which usually are intermittent and at inconsistent doses instead of occurring at a consistent rate and dose magnitude (Nachman et al., 2011). Since epidemiological studies are based on real-world exposures, they provide insight into actual human exposures that can then be linked to diseases, avoiding the uncertainty associated with extrapolation across species. Hence, it can be said that they address the requirements of Regulation 1107/2009 Article 4, which stipulates that the risk assessment should be based on good plant protection practice and realistic use conditions. Thus, epidemiological studies assist problem formulation and hazard/risk characterization whilst avoiding the need for high dose extrapolation (US-EPA 2010).

Epidemiological studies therefore provide the opportunity to a) identify links with specific human health endpoints that are difficult to detect in animal models; b) affirmation of the human relevance of effects identified in animal models; and c) ability to evaluate health effects for which animal models are unavailable or limited (Raffaele et al., 2011). However, in epidemiological studies there are always a variety of factors that may affect the disease outcome and confound the results. For example, when epidemiological data suggest that exposures to pesticide formulations are harmful they usually cannot identify what component may be responsible due to the complexity of accurately assessing human exposures to pesticides. In addition confounding by unmeasured factor(s) associated with the exposure can never be fully excluded. As many diseases are known to be associated with multiple risk factors; a hazard-by-hazard approach is usually considered for evaluating the consequences of individual pesticide hazards on vulnerable systems (Figure 4A). Specifically, single-risk analysis allows a determination of the individual risk arising from one particular hazard and process occurring under specific conditions, while it does not provide an integrated assessment of multiple risks triggered by different environmental stressors (either natural or anthropogenic) (Figure 4B). Risk assessment would benefit by developing procedures for evaluating evidence for co-occurrence of multiple adverse outcomes (Nachman et al., 2011), which is more in line with what happens in human setting. For these reasons, if appropriately conducted, epidemiological studies can be highly relevant for the risk assessment process.

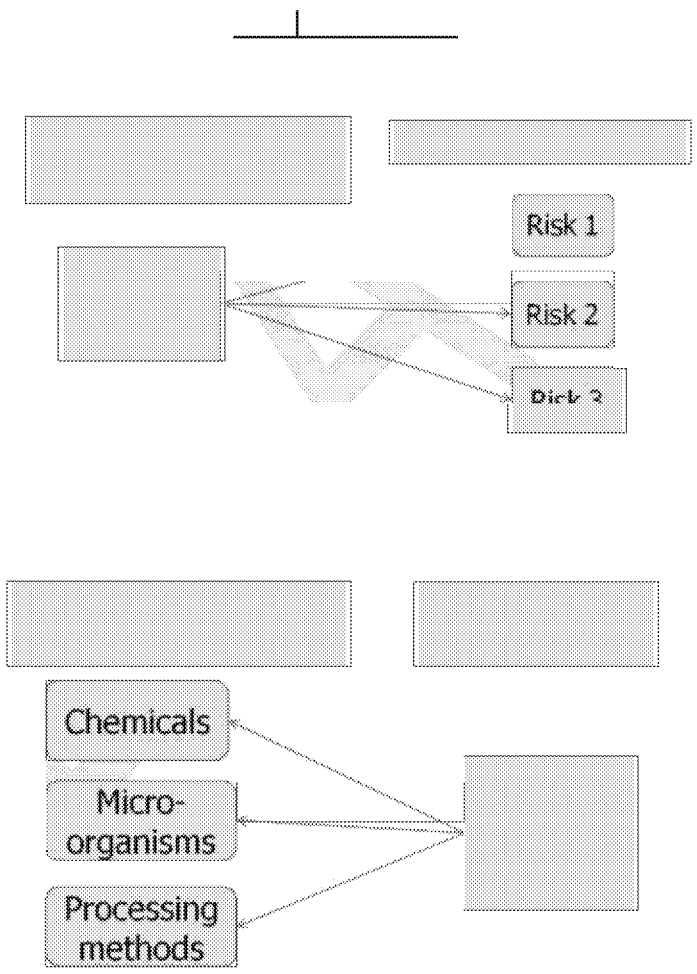

Figure 4: Role of epidemiological studies when compared to classical toxicological studies.

In parallel with epidemiological data, vigilance data can provide an additional stream of evidence, especially for acute toxicity. Cases are usually well-documented and information can be used at different steps of the risk assessment; these include: level and duration of exposure, clinical course and assessment of the causal relationship. In severe cases, the toxin and/or the metabolites are

usually measured in blood or urine which allows for comparison with animal data and in some cases for setting toxicological values.

In summary, experimental studies or epidemiological studies and vigilance data represent two different approaches to collect and assess evidence i.e. one emanating from controlled exposures (usually to a single substance) using experimental study design and a relatively homogeneous surrogate population, the other reflecting the changes observed in a heterogeneous target population from mixed (and varying) exposure conditions using non-experimental study design (ECETOC, 2009). This makes both streams of evidence complementary.

7.2. Principles for weighting of human observational and laboratory animal experimental data

Following the identification of reliable human (epidemiological or vigilance) studies and the assessment of the relevance of the pooled human studies, the separate lines of evidence that were found to be relevant need to be integrated with other lines of evidence that were equally found to be relevant.

The first consideration is thus how well the health outcome under consideration is covered by toxicological and epidemiological studies. When both animal and human studies are considered to be available for a given outcome/endpoint, this means that individual studies will first have been assessed for reliability and strength of evidence (sections 6.2 and 6.3 for epidemiological studies, respectively) prior to the weighting of the various sources of evidence. Although the different sets of data can be complementary and confirmatory, individually they may be insufficient and pose challenges for characterizing properly human health risks. Where good observational data are lacking, experimental data have to be used. Conversely, when no experimental data is available, or the existing experimental data were found not to be relevant to humans, the risk assessment may have to rely on the available and adequate observational studies.

A simple method is proposed for weighting human and experimental studies in order to incorporate them into risk assessment (Figure 5). For a comparative interpretation of human and animal data, this framework should rely on the following principles (adapted from ECETOC, 2009; Lavelle et al., 2012):

Although the totality of evidence should be assessed, only the studies that are found to be reliable (those categorised as acceptable or supplementary evidence) are considered further. If the data from the human or the experimental studies is considered to be of low reliability (categorised as unacceptable), no risk assessment can be conducted.

A WoE approach should be followed where several lines of evidence are found to be relevant. For pesticide active substances, experimental studies following OECD test guidelines are deemed high reliability unless there is evidence to the contrary. The strength of evidence from animal studies can be upgraded if there is high confidence in alternative pesticide toxicity testing methods (e.g., *in vitro* and in *silico* studies). As for epidemiological evidence, the conduct of meta-analysis provides a more precise estimate of the magnitude of the effect than individual studies and also allows for examining variability across studies (see section 6.3).

Next, the studies that are found to be more relevant for the stage being assessed are to be given more weight, regardless of whether the data comes from human or animal studies. Where human data are of highest relevance, they should take precedence for each stage of the risk assessment. When human and experimental data are of equal or similar relevance, it is important to assess their concordance (consistency across the lines of evidence) in order to determine whether and which dataset may be given precedence.

1 In case of concordance between human and animal data, the risk assessment should use all the data as both yield similar results in either hazard identification (e.g. both indicate the same hazard) or hazard characterisation (e.g. both suggest similar safe